記述式答案自動採点における確信度推定とその役割

舟山弘晃 [1,2]　佐藤汰亮 [1,2]　松林優一郎 [1,2]　水本智也 [3,2]　鈴木潤 [1,2]　乾健太郎 [1,2]
[1] 東北大学　[2] 理化学研究所　[3] フューチャー株式会社
{h.funa, tasuku.sato.p6}@dc.tohoku.ac.jp
{y.m, inui, jun.suzuki}@tohoku.ac.jp t.mizumoto.yb@future.co.jp

概要

記述式答案の自動採点は，教育現場における採点コストの低減や公平な採点の実現を目指して研究が行われてきたが，自動採点システムの採点誤りへの懸念が教育現場での盛んな実運用を妨げている．本研究では，自動採点システムの信頼性の担保を目指し，予測の信頼性を表す確信度を導入することで自動採点モデルによる採点誤りを所望の範囲内に抑えるためのタスク設計を提示する．複数の確信度推定手法をテストケースとして前述のタスク設計に基づき実験を行い，目標値の採点誤りの範囲内で自動採点できる可能性を示した．

1　はじめに

記述式問題の自動採点は，事前に人手で作成された採点基準について，入力された文章が採点基準を満たしているか評価し点数として出力するタスクである．主に，大規模な試験において公平かつ低コストな採点を提供するための採点支援や，教育現場における学習支援を目的として古くから研究されてきた [1, 2, 3, 4]．深層学習に基づく自動採点モデルの登場により，近年自動採点システムの性能が向上しており [3, 4, 5]，教育現場での応用への機運が高まっている．

一方で，自動採点システムによる採点誤りの可能性を完全に取り除くことはできない．さらに，そのような採点誤りは学習の妨げになる可能性が指摘されている [6]．舟山ら [7] はこの問題に対して，許容できないような致命的な採点誤り（CSE) という概念を導入し，予測の信頼性を表す確信度を用いて信頼性の低い予測をフィルタリングし，CSE をできる限り取り除ける範囲で自動採点を行い，残りの答案に対する採点は人が行う事を想定し自動採点タスクの再設計を行った．

本研究では，自動採点システムと人間の系において目標の採点精度を実現するという観点から舟山ら [7] のタスク設計を一般の採点誤りに拡張した枠組みを提示する．実験では，この枠組みの下で所望の採点誤りの範囲内に抑えつつ自動採点を実現可能かどうか，テストケースとして複数の確信度推定手法を用いてシュミレーションを行った．その結果，目標採点誤差の設定によっては採点誤りの範囲内で自動採点ができる可能性を示した．

2　確信度を用いた自動採点タスク

自動採点システムの教育現場における実運用に向けて，その採点品質を担保することは重要な課題である．この課題の解決に向けて，我々は新たな確信度を用いた自動採点のスキームとその評価法を提示する．図 1 にその概要を示す．このスキームでは，それぞれの問題や自動採点を導入する教育現場の要請から事前に定められる採点誤差の許容限界があり，このためにシステムによる誤差が一定値を超えるようであればその部分は人手で採点する，という考え方を採用する．そこで，自動採点結果のフィルタリングを行う現実的な手順として，自動採点システムの採点誤差を特定の目標値 e 以下に抑えることができる閾値 τ を開発データで推定し，その閾値を用いて未知の答案に対して自動採点と人手採点の振り分けを行う．

記述式答案の自動採点　まず，記述式答案の自動採点について説明する．本研究では，得点を得るために答案が含むべき情報が明確に定義されるような問題を扱う (Short Answer Scoring: SAS)．各問題に対して予め配点 N とその採点基準が与えられており，採点作業は，生徒が記述した答案に対してこの採点基準に従い自然数の点数を付与するものである．このため，自動採点タスクは，配点 N が定められているそれぞれの問題に対し，入力の答案 $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$ を受け取って答案の点数 $\hat{s} \in S = \{0, \ldots, N\}$ を出力するタスクとして定式化される．

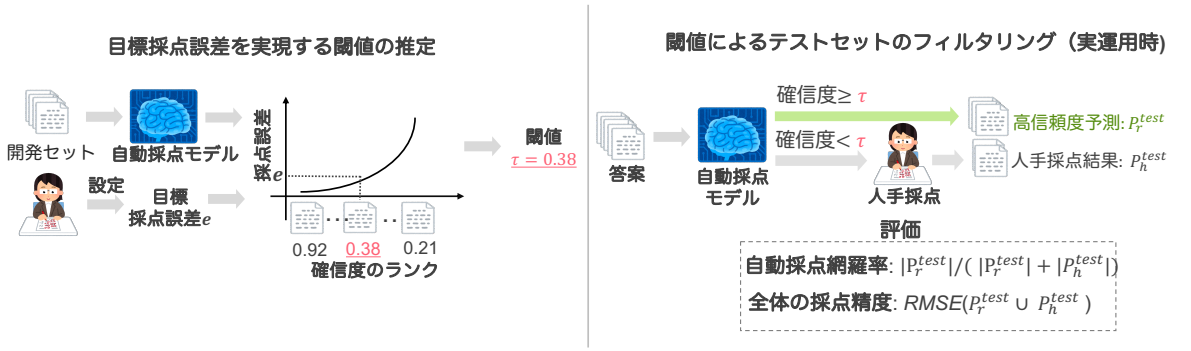

図 1　確信度を用いた自動採点タスクの概要

確信度の導入　ある訓練済みの採点モデル m に関して，このモデルが入力 $\mathbf{x}$ に対し得点 $\hat{s}$ を予測したときの確信度を求める手法 conf があり，この確信度を次のように書くこととする．

$$C_{\mathsf{conf}}(\mathbf{x}, \hat{s}; m) \tag{1}$$

今，l 個のユニークな採点対象の答案 $\{\mathbf{x}_1, ..., \mathbf{x}_l\}$ があり，与えられた答案全てに対する採点結果を $P = \{(\mathbf{x}_i, \hat{s}_i)\}_{i=1}^{l}$ とするとき，このうちの信頼できる採点結果 $P_r \subseteq P$ を次のように求める．

$$P_r = \{(\mathbf{x}, \hat{s}) \in P | C(\mathbf{x}, \hat{s}) \geq \tau\}, \tag{2}$$

ここで，τ はフィルタリングに用いる確信度の閾値である．本研究では，フィルタリングの結果，採点結果が信頼できないと判断された答案の集合 $\overline{P}_r = \{p \in P | p \notin P_r\}$ は十分に訓練された人間が改めて採点することとし，その採点結果には一切の間違いがない理想的な状況を仮定する．$\overline{P}_r$ に対し，人間の採点者が改めて採点した結果を P_h で表す．最終的に，確信度を用いて自動採点と人手採点を振り分けて得られる採点結果 P_f は $P_r \cup P_h$ と表せる．

閾値の推定とフィルタリング　与えられた許容可能な採点誤差 e に対し，開発データに対する採点結果 P^{dev} を用いて確信度の閾値 τ を以下のように求める．

$$\begin{aligned} \underset{\tau}{\text{maximize}} \quad & |P_r^{\mathsf{dev}}| \\ \text{subject to} \quad & \mathsf{Err}(P_f^{\mathsf{dev}}) \leq e \end{aligned} \tag{3}$$

ただし，$\mathsf{Err}(P_r^{\mathsf{dev}})$ は $P_r^{\mathsf{dev}} = \{(\mathbf{x}, \hat{s}) \in P^{\mathsf{dev}} | C(\mathbf{x}, \hat{s}) \geq \tau\}$ 内で発生する採点誤差の総和を表す．次に，開発データにおいて求めた τ を用いてテストセットのフィルタリングを行う．

$$P_r^{\mathsf{test}} = \{(\mathbf{x}, \hat{s}) \in P^{\mathsf{test}} | C(\mathbf{x}, \hat{s}) \geq \tau\}, \tag{4}$$

評価　フィルタリングの結果得られる $|P_r^{\mathsf{test}}|/|P_f^{\mathsf{test}}|$ が自動採点可能な答案の割合（自動採点網羅率）を表し，$\mathsf{Err}(P_f^{\mathsf{test}})$ が答案全体の採点誤差を表す．実験では，答案全体の採点誤差を目標誤差以下に抑えることが可能か，目標値を達成した時の自動採点網羅率はどれくらいか，という 2 点から評価を行う．

3　実験

本研究では，2 節で提示した枠組みにのっとり実験を行い，一般的な採点モデル，確信度推定手法の下で実現可能な採点品質を明らかにするとともに今後の課題を議論する．

3.1　採点モデル

本研究では BERT [8] をベースとした分類モデルを用いる．モデルはまず，入力答案 $\mathbf{x}$ をエンコーダー $\mathsf{enc}(\cdot)$ によって特徴表現 $\mathbf{h} \in \mathbb{R}^{d_h}$ に変換する．

$$\mathbf{h} = \mathsf{enc}(\mathbf{x}) \tag{5}$$

本研究では，$\mathsf{enc}(\mathbf{x})$ の出力として $\mathbf{x}$ を BERT [8] に入力した際に CLS トークンに付与されるベクトルを採用した．ここで d_h は CLS トークンのベクトルの次元数である．その後，得点 $\hat{s}$ を次式により得る．

$$\begin{aligned} p(s|\mathbf{x}) &= \mathsf{softmax}(W\mathbf{h} + b) \\ \hat{s} &= \underset{s \in S}{\mathrm{argmax}}\, p(s|\mathbf{x}) \end{aligned} \tag{6}$$

3.2　確信度推定手法

確信度推定手法として以下の 3 つをテストケースとして考える．

予測確率　本研究で用いる分類モデルにおいて確信度を推定する最も素朴な方法としては，(6) 式で計算される予測確率（Posterior）の値を確信度として考えることができる．

$$C_{\mathsf{prob}}(\mathbf{x}, \hat{s}) = \max_{s} p(s|\mathbf{x}) \tag{7}$$

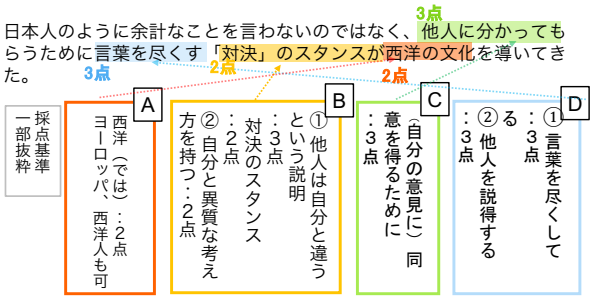

図 2　代々木ゼミナール国語長文読解データセットより答案及び採点基準の例

Trust score　舟山ら [7] は記述式答案自動採点における確信度の推定において Jiang らが提唱した *trust score* [9] を用い，一部の問題でその確信度としての信頼性が予測確率を用いる場合を上回ることを示した．本研究でも *trust score* を確信度推定手法の一つとして考える．本研究では *trust score* を次のように算出する．まず，予め学習データ $\{(\mathbf{x}_1, s_1), \ldots, (\mathbf{x}_k, s_k)\}$ をそれぞれ自動採点モデルに入力し，各入力に対応する特徴表現の集合 $H = \{\mathbf{h}_1, \ldots, \mathbf{h}_k\}$ を得ておく．さらに，これらを得点ラベル s ごとにクラスタ $H_s = \{\mathbf{h}_i \in H | s_i = s\}$ として保持しておく．その上で，未知の入力 $\mathbf{x}$ に対し，その特徴表現 $\mathbf{h}$ と予測得点 $\hat{s}$ を得る．この採点結果 $(\mathbf{x}, \hat{s})$ に関する *trust score* :$C_{\mathrm{trust}}(\mathbf{x}, \hat{s}; m, H)$ は以下を式で算出する．

$$C_{\mathrm{trust}}(\mathbf{x}, \hat{s}; m, H) = \frac{d_c(\mathbf{x},\ H)}{d_p(\mathbf{x},\ H) + d_c(\mathbf{x},\ H)} \tag{8}$$

ただし，

$$d_p(\mathbf{x},\ H) = \min_{\mathbf{h}' \in H_{\hat{s}}} d(\mathbf{h},\ \mathbf{h}'), \tag{9}$$

$$d_c(\mathbf{x},\ H) = \min_{\mathbf{h}' \in (H \setminus H_{\hat{s}})} d(\mathbf{h},\ \mathbf{h}'), \tag{10}$$

であり，$d(\mathbf{h},\ \mathbf{h}')$ は $\mathbf{h}$ から $\mathbf{h}'$ へのユークリッド距離を表す．また，閾値を用いたフィルタリング時に扱いやすいように，値を 0 から 1 の範囲に正規化する目的で新たに分母に d_c を加えている．正規化しても事例間での確信度の大小関係は保持されるためフィルタリングの手順への影響はない．

ガウス過程回帰　一般に記述式答案の自動採点は分類問題及び回帰問題の両方の設定で解くことが可能だが，回帰モデルを用いた場合の方がより高い精度で採点可能であることが示されている [10, 5]．そこで，本研究では確信度を推定可能な回帰モデルの例としてガウス過程回帰 [11] を用いる．本研究では実装として GPytorch [12][1] を用い，3.1 節の分類モデルを学習後，そのエンコーダーが出力する特徴表現を用いてガウス過程回帰 f の学習を行った．未知の入力 $\mathbf{x}$ に対して，この新たな自動採点モデルの予測得点 $\hat{s}$ の確信度 C_{gp} は，$\mathbf{x}$ に対する特徴表現 $\mathbf{h}$ を用いて次のように書ける．

$$\hat{s} = \mathbb{E}[f(\mathbf{h})|H, S], \tag{11}$$

$$C_{\mathrm{gp}} = -\mathbf{Var}[f(\mathbf{h})|H, S] \tag{12}$$

ただし，H は学習データを 5 式に入力して得られた特徴表現の集合，S はその得点の集合である．

3.3　データセット

本研究では，我々が国立情報学研究所（NII）にて公開している代々木ゼミナールの国語長文読解問題データセットを用いる[2]．このデータセットは，各受験者の答案文と採点者によって付与された点数のペアのデータで構成される．図 2 に本研究で対象とする記述問題の具体例を示す．本データセットでは，各問題に対して複数の採点項目が存在し，各項目はそれぞれの採点項目に基づいて独立に人手で採点が行われ，項目点が付与されている．例えば，図 2 では，A，B，C，D の 4 つの採点項目が存在し，それぞれ 2 点，2 点，3 点，3 点が項目点として付与されている．本実験ではこれらの採点項目をそれぞれ独立の問題として扱い，各項目ごとに 3.1 節で述べた採点モデルを訓練し，点数を予測し評価を行う．本研究では 2022 年 1 月時点の版として公開されている 12 問（計 37 項目）を用いた．

3.4　実験設定

3.1 節で述べたように，自動採点モデルのエンコーダーとして事前学習済み BERT [8] を使用し，CLS トークンに付与される値を答案の特徴ベクトルとして採用した[3]．参考のため実験で使用したモデルの採点精度を付録 A に示す．データは学習用，テスト用としてそれぞれ 250 件，250 件に分割を行った．学習用データは更に 5 分割を行い 4 セットを訓練用データ（200 件），1 セットを開発用データ（50 件）とし，学習用のデータを 5 セット作成した．学習に用いたハイパーパラメーターの詳細は付録 B に示す．また，採点誤差を表す関数 Err として，Root

1）https://gpytorch.ai/
2）https://www.nii.ac.jp/dsc/idr/rdata/RIKEN-SAA/
3）事前学習済み BERT は以下の URL の物を使用した：https://github.com/cl-tohoku/bert-japanese

表 1 Posterior（Post.），Trustscore（Trust.），ガウス過程回帰（GP）のそれぞれについて，評価セットで目標値の Root Mean Square Error (RMSE) を達成できるような確信度の閾値を推定し，テストセットのフィルタリングを行った時の RMSE（上段）と自動採点網羅率 [%]（下段）の変化．値は全問題の平均値を表し，±記号は全問題の標準偏差を表す．

目標誤差	0.00	0.02	0.04	0.06	0.08	0.1
Post.	0.026 ± 0.024 23.3 ± 17.3	0.029 ± 0.026 27.4 ± 17.7	0.038 ± 0.030 33.8 ± 18.7	0.054 ± 0.033 44.1 ± 18.1	0.075 ± 0.039 57.7 ± 16.5	0.101 ± 0.042 69.6 ± 14.7
Trust.	0.027 ± 0.021 32.6 ± 12.2	0.030 ± 0.023 36.4 ± 12.9	0.041 ± 0.026 43.8 ± 14.1	0.054 ± 0.029 52.3 ± 13.2	0.077 ± 0.031 63.7 ± 11.7	0.096 ± 0.034 70.9 ± 10.8
GP	0.024 ± 0.022 28.8 ± 14.5	0.026 ± 0.024 31.8 ± 15.1	0.038 ± 0.030 40.2 ± 17.1	0.052 ± 0.038 49.3 ± 19.6	0.075 ± 0.049 61.0 ± 21.2	0.091 ± 0.048 68.3 ± 19.3

Mean Square Error (RMSE) を採用する．また，少量の開発セットの下で，実験結果を安定させるために開発用データ 5 セットを統合した 250 件のデータを用いて閾値を推定した (付録 C).

3.5 結果

この実験では 2 節で述べた手順によって，目標とする採点精度を達成できるような確信度の閾値 τ を検証セットから推定し，その閾値を用いてテストセットに対する予測のフィルタリングを行う．実験では，フィルタリングを行った時に自動採点を行った答案の割合（自動採点網羅率）と，人手採点に一切間違いがなかったと仮定した時の全体の採点精度を求めた．結果は表 1 である．目標とする採点精度が非常に高い時（0.00 〜 0.02）は，確信度によるフィルタリングによって採点誤りを取り除き目標とする採点品質を安定的に実現することは困難であることが示された．一方，許容可能な採点誤差をさらに緩く設定すると（0.04〜），検証セットで決めた閾値を用いてフィルタリングを行うことで目標の採点精度を実現することが可能であることが示唆される．また，いずれの手法も，同目標誤差の下でのフィルタリング後の RMSE の値はおおむね同程度であるが，手法間で自動採点網羅率には大きな差があることが読み取れる．また，採点誤差および自動採点網羅率ともに標準偏差の値は大きく，問題ごとに結果に差があることが示されている．

3.6 議論

実験では一定量の学習データを用いてモデルを訓練し，推論結果のうち信頼性の低かった答案を人が採点することを想定した．しかし，現実の採点現場においては採点にかけられるコストは一定であるため，学習用データの人手採点コストと信頼性の低い答案の人手採点コストはトレードオフの関係にある．すなわち，学習データを増やすとモデルの採点精度は向上し，各得点予測に対する信頼性は全般に向上するが，一方で低信頼度の答案の採点にかけられるコストはその分減少するためより多くの答案を自動採点しなければならない．このように人を含めた系における全体の採点品質を最大化しつつ採点コストを最小化するためには，学習用データの作成にかかる採点コストと，信頼性の低い答案の採点にかけるコストのバランスを最適化する必要がある．この問題の解決に向けて，今後はアクティブラーニングの導入による human-in-the-loop 型の自動採点を想定することで学習データの作成コストと低信頼予測の再採点コストを動的に決定する方法の構築を進めていくことを予定している．

4 おわりに

近年，自動採点システムの採点精度は大きく向上しているが，教育現場へのさらなる応用を進めるためには信頼性の向上が不可欠である．本研究では，舟山ら [7] が提示した確信度を用いた自動採点タスクを拡張し，所望の採点誤りの範囲内で自動採点を行う枠組みを提示することで，採点システムの信頼性の向上に取り組んだ．実験では，既存の確信度推定手法と自動採点モデルを組み合わせることで，目標値の範囲内に採点誤りを抑えつつ自動採点を行える可能性を示した．今後は，アクティブラーニングの活用により学習データの採点コストと信頼性の低い答案の採点コストのバランスを最適化することで，自動採点システムのさらなる信頼性の向上を図りつつ採点コストの最小化を目指す．

謝辞

本研究は，科研費 JP19K12112 の助成を受けたものです．また，実際の模試データを提供していただいた学校法人高宮学園代々木ゼミナールに感謝します．

参考文献

[1] Peter Foltz, Darrell Laham, and T. Landauer. The Intelligent Essay Assessor: Applications to Educational Technology. **Interactive Multimedia Electronic Journal of Computer-Enhanced Learning**, Vol. 1, No. 2, pp. 939–944, 1999.

[2] Yigal Attali and Jill Burstein. Automated Essay Scoring with E-rater v.2.0. **Journal of Technology, Learning, and Assessment**, Vol. 4, No. 3, p. 31, 2006.

[3] Tomoya Mizumoto, Hiroki Ouchi, Yoriko Isobe, Paul Reisert, Ryo Nagata, Satoshi Sekine, and Kentaro Inui. Analytic Score Prediction and Justification Identification in Automated Short Answer Scoring. In **Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications**, pp. 316–325, 2019.

[4] Kaveh Taghipour and Hwee Tou Ng. A Neural Approach to Automated Essay Scoring. In **Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing**, pp. 1882–1891, November 2016.

[5] Brian Riordan, Andrea Horbach, Aoife Cahill, Torsten Zesch, and Chong Min Lee. Investigating neural architectures for short answer scoring. In **Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications**, pp. 159–168, 2017.

[6] Oleg Sychev, Anton Anikin, and Artem Prokudin. Automatic grading and hinting in open-ended text questions. **Cognitive Systems Research**, Vol. 59, pp. 264–272, 2020.

[7] Hiroaki Funayama, Shota Sasaki, Yuichiroh Matsubayashi, Tomoya Mizumoto, Jun Suzuki, Masato Mita, and Kentaro Inui. Preventing critical scoring errors in short answer scoring with confidence estimation. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop**, pp. 237–243, Online, July 2020. Association for Computational Linguistics.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 4171–4186, June 2019.

[9] Heinrich Jiang, Been Kim, Melody Y. Guan, and Maya R. Gupta. To Trust Or Not To Trust A Classifier. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, **Proceedings of Advances in Neural Information Processing Systems 31**, pp. 5546–5557, 2018.

[10] Stig Johan Berggren, Taraka Rama, and Lilja Øvrelid. Regression or classification? automated essay scoring for Norwegian. In **Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications**, pp. 92–102, Florence, Italy, August 2019. Association for Computational Linguistics.

[11] Carl Edward Rasmussen. **Gaussian Processes in Machine Learning**, pp. 63–71. Springer Berlin Heidelberg, Berlin, Heidelberg, 2004.

[12] Jacob R. Gardner, Geoff Pleiss, David Bindel, Kilian Q. Weinberger, and Andrew Gordon Wilson. Gpytorch: Blackbox matrix-matrix gaussian process inference with gpu acceleration, 2021.

A　実験に用いたモデルの採点精度

実験に用いた自動採点モデルの Quadratic Weighted Kappa (QWK) と RMSE を示す．値は各問題ごとに，その問題の全項目を平均したものを示す．

表 2　分類モデルとガウス過程回帰の QWK

	分類モデル	ガウス過程回帰
Y14_1-2_1_3	0.943	0.944
Y14_1-2_2_4	0.882	0.891
Y14_2-1_1_5	0.767	0.782
Y14_2-1_2_3	0.570	0.554
Y14_2-2_1_4	0.836	0.850
Y14_2-2_2_3	0.773	0.795
Y15_1-1_1_4	0.748	0.766
Y15_1-3_1_2	0.855	0.866
Y15_1-3_2_4	0.228	0.213
Y15_2-2_1_5	0.867	0.872
Y15_2-2_2_4	0.701	0.750
Y15_2-2_2_5	0.694	0.735
Avg.	0.739	0.752

表 3　分類モデルとガウス過程回帰の RMSE

	分類モデル	ガウス過程回帰
Y14_1-2_1_3	0.083	0.079
Y14_1-2_2_4	0.136	0.126
Y14_2-1_1_5	0.161	0.141
Y14_2-1_2_3	0.327	0.301
Y14_2-2_1_4	0.159	0.143
Y14_2-2_2_3	0.266	0.240
Y15_1-1_1_4	0.277	0.247
Y15_1-3_1_2	0.151	0.140
Y15_1-3_2_4	0.276	0.260
Y15_2-2_1_5	0.158	0.146
Y15_2-2_2_4	0.172	0.148
Y15_2-2_2_5	0.278	0.244
Avg.	0.204	0.185

B　ハイパーパラメータ

分類モデルとガウス過程回帰の学習に用いたハイパーパラメータは次の通りである．

表 4　分類モデル学習に用いたハイパーパラメータ

最適化アルゴリズム	Adam
学習率	1.00E-05
バッチサイズ	16
最大エポック数	30

表 5　ガウス過程回帰の学習に用いたハイパーパラメータ

最適化アルゴリズム	Adam
学習率	1.00E-01
最大エポック数	30

C　少量データにおける閾値の推定

2 節で述べたように，開発セットを用いて閾値の推定を行う必要がある．開発セットのサイズが大きいほど安定的に閾値を決めることが可能であると考えられるが，答案の採点コストを考えると実際の教育現場において開発セットに用いることができるデータの量は多くはないことが想定される．本実験においても，各５０件開発証データから目標を達成するための閾値を安定的に推定することは難しい．そこで，本実験では検証用データ 5 セットを統合し 250 件としたものを用いて閾値の推定を行った．その閾値を用いて各５セットの学習データで訓練したモデルのテストデータに対する採点結果のフィルタリングを行った．